Sales Performance of Video Games

Abstract

The gaming industry has accumulated a wealth of data over many years, encompassing gamers' preferences and behaviors. Game developers can leverage this data to enhance their games. Predicting sales trends is crucial for gaming businesses, yet there's a lack of research on the factors influencing sales prediction. Machine learning techniques offer powerful tools for uncovering insights within vast datasets, improving prediction accuracy and efficiency. This paper provides a brief analysis of gaming sales data and the methods employed for sales prediction.

Introduction

The video game industry needs precise sales data to accommodate its exponential market growth. Over the past decade, there has been a remarkable surge in revenue generated by computer and video games in the United States (Clement 2021). Consequently, there is a necessity to forecast the purchasing behavior of various video game enthusiasts by leveraging historical sales information. This paper focuses on supervised and unsupervised learning. Identifying the best variable that predicts global sales. Also, to identify the variables that best classify and cluster the sales. To achieve this, we have employed machine learning techniques to make predictions and classifications regarding video game sales in the market.

Methodology

The method used in this report contains a supervised and unsupervised model. The overview of each model used is given below:

Linear Regression model

Linear regression is a frequently employed method in predictive modeling. Its primary objective is to determine a mathematical equation that relates continuous variable Y to one or more X variables. This algorithm aims to establish a connection between these variables, with one being the predicted variable and the other being the outcome variable, whose value is determined based on the predictive variable (Gopalakrishnan et al., 2018).

Random Forest

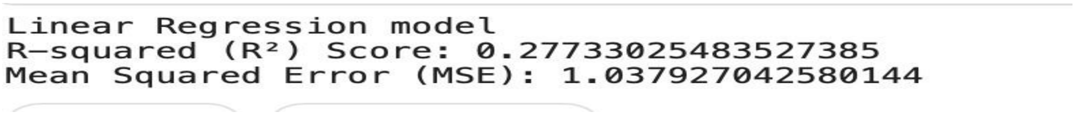
The Random Forest algorithm is a supervised machine-learning technique that constructs a forest comprising multiple trees through a random process (David et al., 2013). While Decision Trees are simple to implement and perform well with training data, they often suffer from lower accuracy, primarily due to a phenomenon known as overfitting. Overfitting arises when a model becomes overly tailored to the training data to the extent that it adversely affects its performance on new, unseen data. This is where Random Forest proves to be a valuable alternative (Boinee et al., 2018).

Result

This section gives the answer to the questions, which is the result of the analysis.

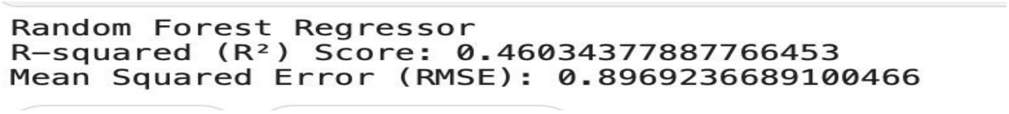
Question 1:

Firstly, all the sales columns of the games were dropped from the dataset except the global sales which is the sum of the other sales from different regions. To check which of the variables best predicts the sales, I first compare linear regression output (Fig. 1) with random forest regression (Fig. 2).



```
Linear Regression model
R–squared (R²) Score: 0.27733025483527385
Mean Squared Error (MSE): 1.037927042580144
```

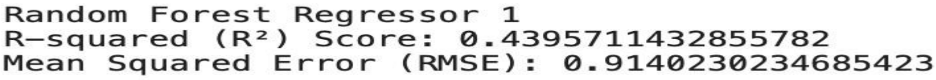
Figure 1: Linear regression



```
Random Forest Regressor
R–squared (R²) Score: 0.46034377887766453
Mean Squared Error (RMSE): 0.8969236689100466
```

Figure 2: random forest regression

The random forest regression had better accuracy than the linear regression. Now, the important features from the random forest regression were computed and the variables with lower impact were dropped and the training was done again. Figure 3 shows the result when the variables with lower impact were dropped.



```
Random Forest Regressor 1
R-squared (R²) Score: 0.4395711432855782
Mean Squared Error (RMSE): 0.9140230234685423
```

Figure 3: random forest regression

In conclusion, the combination of the variables in the video game dataset that best predicts global sales (excluding the other sales) using Random Forest regression is Platform, Year of Release, Genre, Publisher, Critic Score, Critic Count, User Score, User Count, Developer, Rating.

Question 2:

Using the Random Forest regressor, the graph below shows the feature importance which shows the contribution of variables to the target.

In North America (figure 4) shows the contribution and the effect of user count, user score, critic count, and critic score on the sales of video games. It can be seen that the user count has the highest contribution followed by the critic count which user score contributes the least among the four variables.

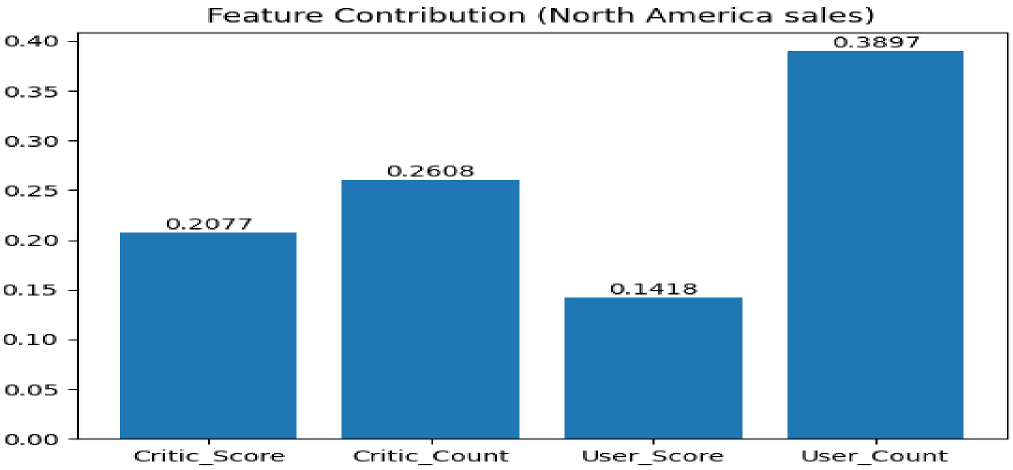


Figure 4: North American sales

Figure 5 shows the impact of the number of critics and users as well as their review scores on the sales of Video games in Europe. The plot shows that User count had the highest impact on Europe sales, and user score had the least impact on the sales of Video games in Europe.

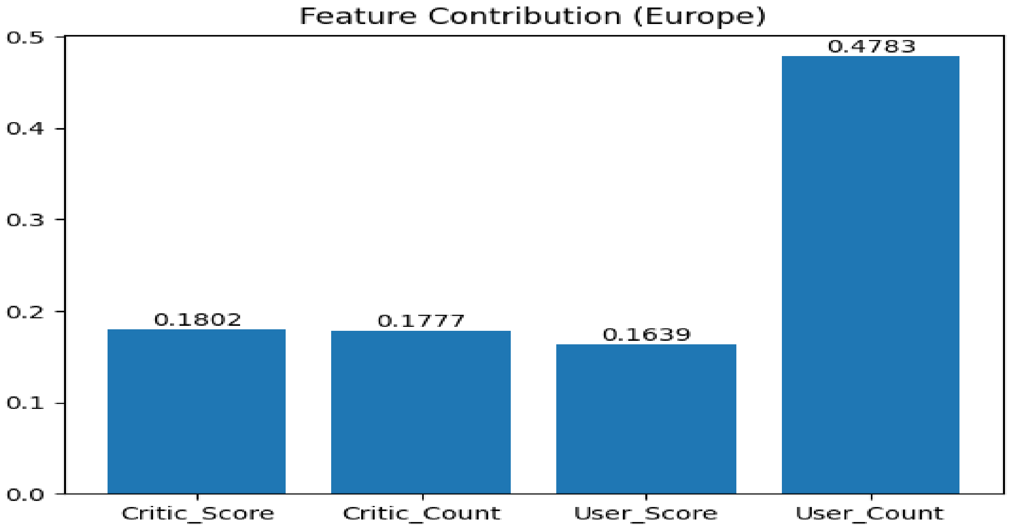


Figure 5: Europe sales

Table 6 shows that the User count has the highest impact on the number of sales in Japan, followed by the Critic count, and the User score had the least impact.

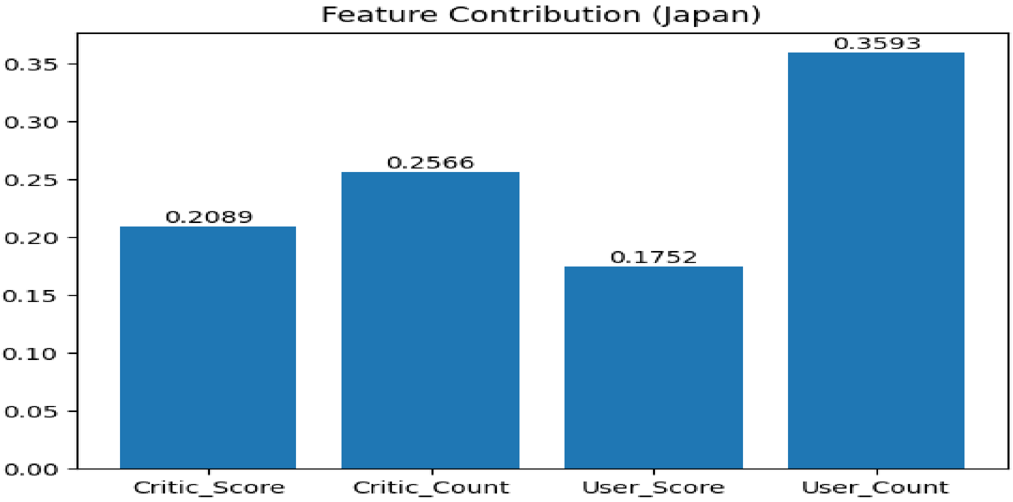
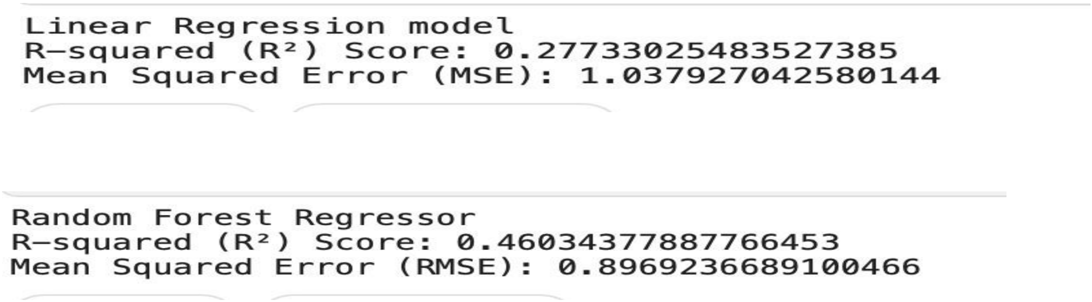


Figure 6: Japan sales

Question 3:

The random forest regressor was used for this task because it gave a lower RMSE compared to the baseline model linear regression analysis. Also, it have the highest R-square value (figure 7).



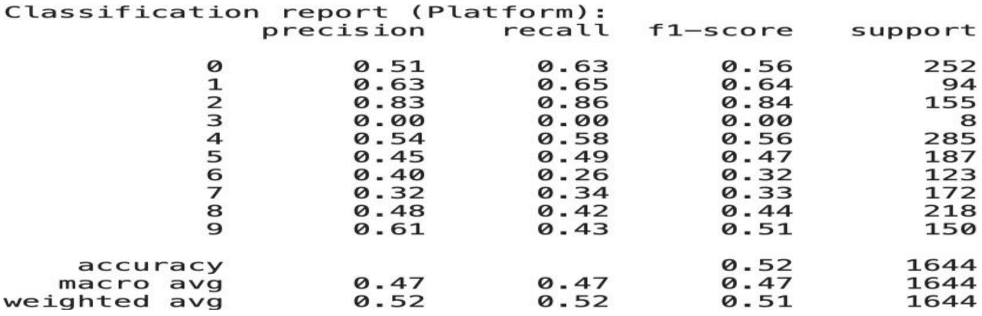
```
Linear Regression model
R-squared (R²) Score: 0.27733025483527385
Mean Squared Error (MSE): 1.037927042580144
```

```
Random Forest Regressor
R-squared (R²) Score: 0.46034377887766453
Mean Squared Error (RMSE): 0.8969236689100466
```

Figure 7: R-square and RMS

Question 4:

From the classification report, using the f1-score weighted average because of the imbalance in the dataset, it can be seen that “Rating” is the variable that performed best in classifying the dataset with an accuracy of 70%.

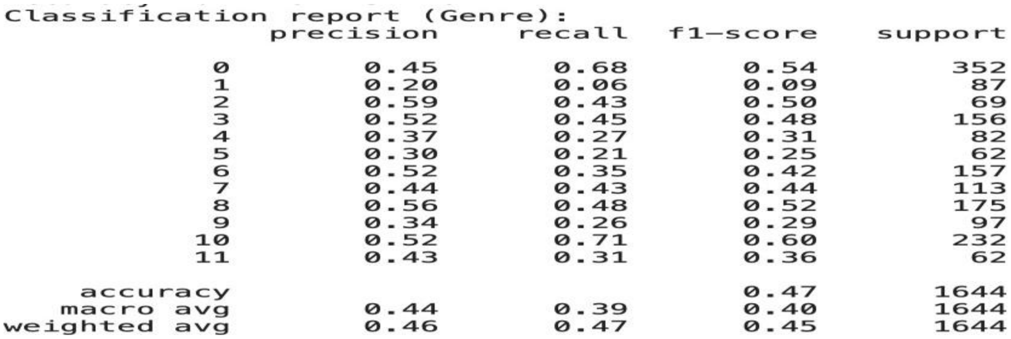


```
Classification report (Platform):
              precision    recall  f1-score   support

           0       0.51      0.63      0.56       252
           1       0.63      0.65      0.64        94
           2       0.83      0.86      0.84       155
           3       0.00      0.00      0.00         8
           4       0.54      0.58      0.56       285
           5       0.45      0.49      0.47       187
           6       0.40      0.26      0.32       123
           7       0.32      0.34      0.33       172
           8       0.48      0.42      0.44       218
           9       0.61      0.43      0.51       150

    accuracy                           0.52      1644
   macro avg       0.47      0.47      0.47      1644
weighted avg       0.52      0.52      0.51      1644
```

Figure 8: Platform metrics

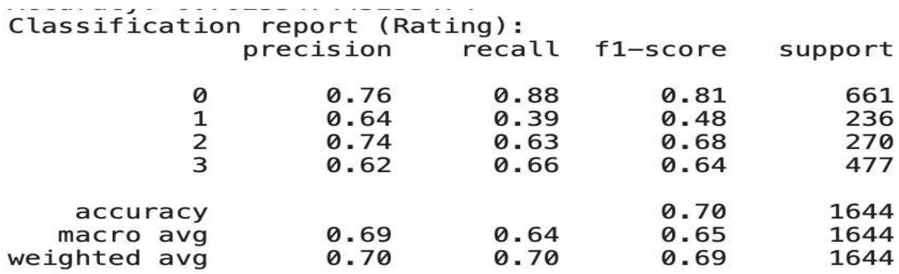


```
Classification report (Genre):
              precision    recall  f1-score   support

           0       0.45      0.68      0.54       352
           1       0.20      0.06      0.09        87
           2       0.59      0.43      0.50        69
           3       0.52      0.45      0.48       156
           4       0.37      0.27      0.31        82
           5       0.30      0.21      0.25        62
           6       0.52      0.35      0.42       157
           7       0.44      0.43      0.44       113
           8       0.56      0.48      0.52       175
           9       0.34      0.26      0.29        97
          10       0.52      0.71      0.60       232
          11       0.43      0.31      0.36        62

    accuracy                           0.47      1644
   macro avg       0.44      0.39      0.40      1644
weighted avg       0.46      0.47      0.45      1644
```

Figure 9: Genre metrics



```
Classification report (Rating):
              precision    recall  f1-score   support

           0       0.76      0.88      0.81       661
           1       0.64      0.39      0.48       236
           2       0.74      0.63      0.68       270
           3       0.62      0.66      0.64       477

    accuracy                           0.70      1644
   macro avg       0.69      0.64      0.65      1644
weighted avg       0.70      0.70      0.69      1644
```

Figure 10: Rating metrics

Question 5:

Cross-validation technique was used to determine if the model overfit or not. Therewas no high variance in cross-validation scores using the macro F1-score for the three classification models.

Question 6:

Based on the performance of the model, it is not ideal for deployment especiallygiven it will give a misleading prediction 30% of the time is quite high.

Question 7:

Table 1 shows the metric for the clustering. It can be seen that the rating variable best describes the group formed using the silhouette score. But when using the score, the platform was seen to be the best that describes the group formed.

Table 1:

Variable	Silhouette score	Adjusted rand score
Platform	0.3669	0.0945
Genre	0.3586	0.0080
Rating	0.3689	0.0850

Conclusion

In conclusion, the paper focused on analysing video game global sales data. It was discovered that the best combination of these variables (Platform, Year of Release, Genre, Publisher, Critic Score, Critic Count, User Score, User Count, Developer, and Rating) best predicts global sales. Also, game rating was seen to be the best in classifying the data set.

References

Clement, J. 2021. Number of video gamers worldwide in 2021, by region. Available at: https://www.statista.com/statistics/293304/number-video-gamers/ [Accessed: 05sep 2023]

Buckley, D., Chen, K., & Knowles, J. (2013). Predicting skill from game play input toa first-person shooter. IEEE. ISBN: 978-1-4673-5311-3/13.

Gopalakrishnan, T., Choudhary, R., & Prasad, S. (2018). Prediction of Sales Value inOnline shopping using Linear Regression. IEEE. DOI: 10.1109/CCAA.2018.8777620.

Boinee, P., Angelis, A. D., & Foresti, G. (2015). Meta random forests. InternationalJournal of Computational Intelligence, 2(3), 138-147.